Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Jakub Náplava		
Název práce	Automatická oprava pravopisu		
Rok odevzdání	2017		
Studijní program	Informatika	**Studijní obor**	Umělá inteligence

Autor posudku	Milan Straka	**Role**	Vedoucí
Pracoviště	Ústav formální a aplikované lingvistiky		

Text posudku:

Diplomová práce se zabývá návrhem algoritmů pro opravu textů v přirozeném jazyce – od jednoduchých oprav jako generování diakritických znamének přes opravu překlepů až k obecné úloze korekce všech gramatických chyb v textu. Oprava pravopisu a gramatických chyb v textu je úloha s četnými aplikacemi, které denně využívá velké množství lidí – pokrok v této oblasti má znatelný přínos.

První dvě kapitoly práce popisují samotnou úlohu opravu textu v přirozeném jazyce včetně několik jejích variant (doplňování diakritiky, oprava překlepů a oprava gramatiky) a existující datasety.

Následující tři kapitoly popisující tři architektury hlubokých neuronových sítí, které je možné použít k řešení postupně obecnějších variant cílové úlohy – a to na úrovni znaků (oprava diakritiky, i/y, doplňování čárek, velká/malá písmena, atd.), dále na úrovni slov (libovolné opravy v rámci jednoho slova) a nakonec na úrovni kompletních vět (umožňující například změnu pořadí všech slov ve větě). Z těchto tří architektur jsou první a třetí existující; architektura pro opravu chyb na úrovni slov je navržena autorem.

Vyhodnocení navržených modelů provádí autor v kapitole šesté. Vyhodnocení je provedeno na čtyřech datasetech, z nichž dva autor připravil a publikoval. Samotná implementace je popsána v kapitole sedmé.

Práci považuji za velmi kvalitní. Autor využívá pokroků hlubokého učení v oblasti zpracování přirozeného textu a navrhuje několik modelů řešící postupně komplikovanější varianty opravy textů. Tyto modely jsou pak podrobně a korektně vyhodnoceny na existujících i nově vytvořených datasetech a porovnány s dostupnými existujícími aplikace včetně nástrojů pro opravu pravopisu v aplikacích Microsoft Word či Google Chrome. Pro češtinu dosahuje autor nejlepší známých výsledků (například navržený model pro generování diakritiky redukuje počet chyb skoro na polovinu oproti nejlepšímu existujícímu řešení), pro angličtinu pak porovnatelných výsledků s existujícími řešeními.

Kromě vysoké přesnosti mají navržené modely podstatnou výhodu v tom, že jsou jazykově nezávislé a učí se pouze z trénovacích dat – je možné je použít na libovolný jazyk, pro který jsou k dispozici texty s chybami a bez nich. To je rozdíl oproti mnoha existujícím systémům, které vyžadují buď jazykově závislé komponenty (například model pro generování kandidátních oprav) a/nebo analýzy (například lemmatizaci či slovní druhy).

Dosažené výsledky je možno spolehlivě reprodukovat, protože vytvořené modely implementované ve frameworku TensorFlow jsou zveřejněny včetně zvolených hyperparametrů, a stejně tak jsou zveřejněny všechny použité datasety.

Práce je napsaná velmi dobrou angličtinou a obsahuje zdařile zpracovanou rešerši současného stavu poznání.

Celkově hodnotím diplomovou práci jako velice povedenou a prokazující schopnost samostatné výzkumné činnosti.

Práci doporučuji **k obhajobě.**

Práci navrhuji **na zvláštní ocenění.**

Pokud práci navrhujete na zvláštní ocenění (cena děkana apod.), prosím uveďte zde stručné zdůvodnění (vzniklé publikace, významnost tématu, inovativnost práce apod.).

Práci považuji za velmi zdařilou. Zabývá se tématem opravy textu v přirozeném jazyce (oprava pravopisu, gramatiky, diakritiky, …) s množstvím praktických aplikací, používá moderní metody hlubokého učení k navržení jazykové nezávislých modelů a dosahuje pro češtinu nejlepších známých výsledků. Úroveň popisu modelů a experimentů, spolu se zveřejněním zdrojových kódů i datasetů, umožňuje spolehlivou replikaci výsledků. V neposlední řadě obsahuje práce kvalitní rešerši existujících řešení.

Datum 1. června 2017 **Podpis**